International Conference on Recent Innovations in Science, Agriculture, Engineering and Management
University College of Computer Applications Guru Kashi University, Bathinda, Punjab (India)
20th November 2017, www.conferenceworld.in ISBN: 978-93-86171-80-1

IMPLEMENTATION AND APPLICATIONS OF BIG DATA IN MEDICAL SCIENCE

Sukhpreet Singh

Research Scholar, M. Phil (Comp. Sc.), Punjabi University, Patiala (PB),(India)

ABSTRACT

Medically related data collection saw a huge increase in the past few decades. This tremendous increase in clinical data is referred to as a big data in medical science. These huge medical datasets brings many challenges related to storage, processing and analysis along with them. In medical science big data is expected to play an important role in identifying causality of patient symptoms, in predicting hazards of disease incidence or reoccurrence and in improving primary health care quality. Big Data can be combined with new technology to bring about positive conversion in the health care segment. A technology aimed at making Big Data analytics a certainty will act as a key element in transforming the way the health care industry operates today. The study and analysis of Big Data can be used for tracking and managing population health care effectively and efficiently In ten years, eighty percent of the work people do in medicine will be replaced by technology. And medicine will not look anything like what it does today. In this research paper we will discuss about the role played by big data in health care segment, technology used to handle big data, challenges faced by big data, obstacles in using big data in the health industry, how big Data analytics can take health care to a new level by enhancing the overall quality of patient care.

Keywords— Big data; framework; medical science; EMR; HIS.

I. INTRODUCTION

Big data [1], [2], [12] are rapidly all over the place. Everyone seems to be collecting, analyzing, and making money from it. No matter whether we are talking about analyzing zillions of Google search queries to predict flu outbreaks, or zillions of phone records to detect signs of terrorist activity, or zillions of airline stats to find the best time to buy plane tickets, big data are on the case. By combining the power of modern computing with the enormous data of the digital era, it promises to solve virtually any problem like crime, public health, the evolution of grammar, etc.

The goal of big data management is to ensure a high level of data quality and accessibility for business intelligence and big data analytics applications. Corporations, government agencies and other organizations employ big data management strategies to help them contend with fast-growing pools of data, typically involving many terabytes or even petabytes of information saved in a variety of file formats. Effective big data management helps companies locate valuable information in large sets of unstructured data and semi-structured data from a variety of sources, including call detail records, system logs and social

media sites. Most big data environments go beyond relational databases and traditional data warehouse platforms to incorporate technologies that are suited to processing and storing non- transactional forms of data. The increasing focus on collecting and analyzing big data is shaping new platforms that combine the traditional data warehouse with big data systems in a logical data warehousing architecture. As part of the process, it must be decided what data must be kept for compliance reasons, what data can be disposed of and what data should be kept and analyzed in order to improve current business processes or provide a business with a competitive advantage. This process requires careful data classification so that ultimately, smaller sets of data can be analyzed quickly and productively.

II. HADOOP AND ITS ARCHITECTURE

Hadoop is a java based framework that is efficient for processing large data sets in a distributed computing environment. Hadoop is sponsored by Apache Software Foundation. The creator of Hadoop was Doug Cutting and he named the framework after his child's stuffed toy elephant. Applications are made run on systems with thousands of nodes making use of thousands of terabytes via Hadoop. Distributed file system in Hadoop facilitates fast data transfer among nodes and allows continuous operations of the system even if node failure occurs. This concept lowers the risk of disastrous system failure even if multiple nodes become inoperative. The inspiration behind working of Hadoop is Google‟ s Map reduce which is a software framework in which application under consideration is broken down into number of small parts. Any of these fragments can run on any node in the cluster. The components involved in Hadoop ecosystem are Hadoop kernel, Map Reduce, Hadoop distributed file system and many related projects like Apache hive, HBase and Zookeeper. The use of Hadoop framework is done by major players like Yahoo, IBM and Google. Ideal operating systems for Hadoop are Windows and Linux, but it can also work with BSD and OS X. The technology involved was developed by Google during their earlier days in order to index all valuable textural and structural information collected by them. All this was done to provide meaningful result to the user. Later this innovation of google was integrated into Nutch which was an open source project and Hadoop was spun off from it. Yahoo was the most prominent in developing Hadoop for enterprise applications. Hadoop platform is preferred to solve problems where data is enormous and does not fit appropriately in tables. It is used to run analytics that involves extensive computation. There are numerous sectors where Hadoop finds it application. It is used in finance to perform exact portfolio evaluation and risk analysis. In online retail, it helps in providing better answers to the customers to increase the probability of their buying the things.

Hadoop runs on large number of machines that do not share any memory space. It is similar to buying a whole bunch of commodity servers, slap them in a rack and run Hadoop software on each one of them. Suppose one decides to load all the data of their organization in Hadoop, the software busts that data into pieces and spread them across different servers. Hadoop keeps track of where the whole data resides. Also mentionable is that if data on server goes offline, it can be replicated automatically from a known good copy. In centralized database systems one big disk that is connected to four, eight or sixteen processors. But in comparison, each of these servers has two, four or eight CPUs in Hadoop. You can run your indexing job by

sending your code to each of the dozens of servers in your cluster, and each server operates on its own little piece of the data. The result is then delivered back collectively. This is what is known as map - reduce. Map reduce technique involves mapping the operations out to all these servers and thereafter these results are reduced back into a single result set.As Map Reduce is an algorithm, it can be written in any programming language. Hadoop map reduce works in three.

First Stage*: *Mapping: In this stage, a list of elements is provided to a „mapper" function to get it transferred into pairs. The mapper function does not modify the input data, but simply returns a new output list.

Intermediate Stages*: *Shuffling and Sorting: After the mapping stage, the program exchanges the intermediate outputs from the mapping stage to different „reducers" . This process is called shuffling.

***Final Stage*: *Reducing*:** In the final reducing stage, an instance of a user-provided code is called for each key in the partition assigned to a reducer. In particular, we have one output file per executed reduce task.

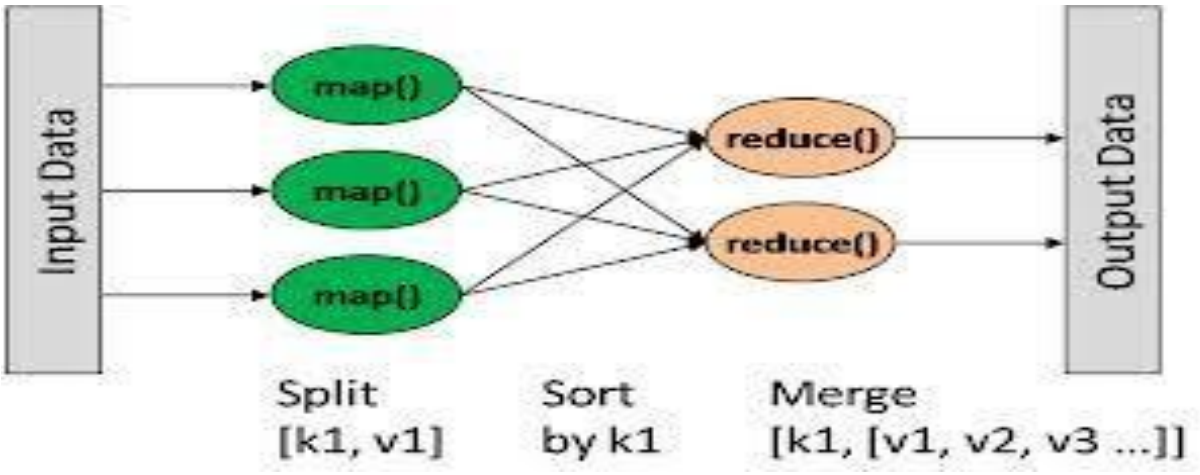

Fig.1. Working of map reduces.

III. ISSUES RELATED WITH BIG DATA CHARACTERISTICS

Data Volume – With the increase in volume, the worth of different data records will decrease in proportion to age, type, richness, and quantity among other factors.

Data Velocity – Data is being generated at tremendous speed with each minute passing. The velocity at which this data is being generated is beyond the handling power of traditional systems.

Data Variety - Mismatched data formats, non-aligned data structures, and inconsistent data semantics represents significant challenges that can lead to analytic collapse.

Data Value – Often it is witnessed that there is a huge gap in between the business leaders and the IT professionals. The main concern of business leaders is to just add value to their business and to maximize their profit. On the other hand, IT leaders deal with technicalities of the storage and processing.

Data Complexity - Data scientists have to link, match, cleanse and transform data across systems coming from various sources. It is also necessary to connect and correlate relationships, hierarchies and multiple data linkages or data can quickly spiral out of control [3].

Data Veracity - Veracity refers to the messiness or trustworthiness of the data. With many forms of big data quality and accuracy are less controllable [4], [10].

IV. PHASES INVOLVED IN BIG DATA

Big data definitely have some source of origin. It is not created from a vacuum. Different scientific experiments being carried out in the world today produces petabytes of data per day. Much of this data is of no use and has to be filtered out. The first challenge faced is to set filtering parameters as such that useful data doesn" t gets discarded. For example, suppose one sensor reading differs substantially from the rest: it is likely to be due to the sensor being faulty, but how can we be sure that it is not an artifact that deserves attention? We need research in the science of data reduction that can intelligently process this raw data to a size that its users can handle while not missing the needle in the haystack.

The second challenge encountered is related to automatically generating right metadata to illustrate what data is recorded, how it is recorded and measured. In scientific experiments, considerable detail regarding specific experimental conditions and procedures may be required to be able to interpret the results correctly, and it is important that such metadata be recorded with observational data.

Information Extraction and Cleaning

It is mention able here that information collected is not in an analysis ready format. For example, consider the collection of electronic health records in a hospital, comprising transcribed dictations from several physicians, structured data from sensors and measurements, and image data such as x- rays. The data in this format cannot be effectively analyzed An information extraction process should be applied to such data to pull out the required information from the sources under consideration and present it in a structured format suitable for analysis. This is really a big challenge. This data may include images and videos and such extraction is highly application dependent.

Data Integration, Aggregation, and Representation

It is not enough to merely collect, record and throw the data into a repository. If we have large data sets in repository, then it will be almost impossible for the user to find the desired data when required. But with sufficient amount of metadata there is some hope but still challenges persists due to differences in experimental details and in data record structure. Data challenging is much more than simply locating,

identifying, understanding and citing data. All this process needs to occur in a complete automated manner for an effective large scale analysis. Suitable database design is most important. There are many different ways in which data can be stored. Certain designs will be better than others for certain purposes and possibly may carry drawbacks for other purposes. Therefore it can be concluded that database design is an art and needs to be carefully executed by trained professionals.

Query Processing, Data Modelling, and Analysis Methods for Querying and Mining

There is no doubt in the fact that big data is noisy, dynamic, diverse, inter-related and untrustworthy. But even noisy Big Data could be more valuable than tiny samples because general statistics obtained from frequent patterns and correlation analysis usually overpower individual fluctuations and often disclose more reliable hidden patterns and knowledge. Further, interconnected Big Data forms large heterogeneous information networks, with which information redundancy can be explored to compensate for missing data, to crosscheck conflicting cases, to validate trustworthy relationships, to disclose inherent clusters, and to uncover hidden relationships and models. Mining requires integrated, cleaned, trustworthy, and efficiently accessible data, declarative query and mining interfaces, scalable mining algorithms, and big-data computing environments. At the same time, data mining itself can also be used to help improve the quality and trustworthiness of the data, understand its semantics, and provide intelligent querying functions. A problem with current Big Data analysis is the lack of coordination between database systems, which host the data and provide SQL querying, with analytics packages that perform various forms of non-SQL processing, such as data mining and statistical analyses.

Interpretation

The analysis of big data remains of no value if users are not able to understand the analysis concept. Decision maker is provided with the result of analysis and is expected to interpret these results. This interpretation requires efforts. It involves deeply examining all the assumptions made and retracing the analysis. There are several sources of errors like system may carry bugs and conclusions may be based on error prone data. No responsible user will yield authority to computer system for all this. Instead one will try to understand and verify the results produced by computer system. All this should be made easy by computer system and this is a big challenge with big data due to its complexity.

V. BIG DATA IN MEDICAL SCIENCE

Devices like cell phones which are a part of our daily life provide us with huge stream of data about human life and behaviour. Behavioural data obtained from these devices when combined with existing health data can greatly open opportunities to predict long term health conditions and design better diagnostic tools, prevent diseases, increase access to health care and reduce cost of health care. Prominent application areas are mental health, environmental health, chronic and infectious disease, health care cost, health care quality, accidents

 ISBN: 978-93-86171-80-1

and injury. Along with theses promises, there are also certain issues related to big data in medical science like data privacy and ownership, theft of personal data and its misuse and new scientific risks.

Fig. 2. Technology involved in health care segment.

Medical care and medically related research is becoming bigger and complex day by day and is resulting in huge volumes of data. All this is further propagated with the advent of new technologies. The aim to advance medical care and convert science into modern science is confined by the ability to process big data efficiently and effectively. The sources of data in medical science ranges from human genetics and pathogen genomics to routine clinical documentation, from internal imaging to motion capture, from digital epidemiology to pharmacokinetics and from treatment pathways to life course assessment [9].

Health care is one of the top social and economic issues in many countries, such as the India, the UK, South Korea, The United States and even middle-income countries. In India, health care sector suffers from underfunding and bad governance. No doubt, India has made huge improvements since independence; majority (70%) of the effort has been led by the private sector. Still India accounts for 21% of the world" s burden of disease[4].

VI. CONCLUSION

The real issue is not that we are acquiring large amounts of data. It's what you do with the data that counts. Today, a significant proportion of the cost and time spent in the drug development process is attributable to unsuccessful formulations. By enabling researchers to identify compounds with a higher likelihood of success, Big Data can help reduce the cost and the time to market for new drugs. Also, by integrating learning from medical data in the early stages of development, researchers will now be able to customize drugs to suit aggregated patient profiles [6].

Currently, information privacy concerns are the single biggest obstacle to Big Data adoption in health care. Another is the absence of an analytics solution powerful enough to gather massive volumes of largely unstructured health data, perform complex analyses quickly, and trigger meaningful solution, for instance, gather all the data from ICU monitors, which today goes un-stored, put it on the Cloud, decipher significant medical patterns that are yet undiscovered, and trigger a medical action instead of merely an alarm.

By providing an overview of the current state of big data applications in the healthcare environment, this research paper has explored the existing challenges that governments and healthcare stakeholders are facing. All big data projects in leading countries and healthcare industries have similar general common goals, such as the provision of easy and equal access to public services, better citizens' healthcare services, and the improvement of medical-related concerns. However, each government or healthcare stakeholder has its own priorities, opportunities, and threats, based on its country's unique environment (e.g., healthcare expenditures in the United States, the inefficient and wasteful healthcare system in Japan, regional disparities in the healthcare resources in India, etc.) which big data projects must address.

REFERENCES

[1] E. Dumbill, "Big data market survey: Hadoop solutions," Radar Insight, Analysis, and Research About Emerging Technologies, 2012. [2] www.forbes.com/sites/lisaarthur/2013/08/15/what-is-big-data/

[3] http://www.business2community.com/digital-marketing/4-vs-big-data- digital-marketing-0914845#!bgCHyQ

[4] http://dashburst.com/infographic/big-data-volume-variety-velocity/

[6] http://www.informationweek.in/informationweek/cio-blog/175124/health care-compulsion-choice

[7] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3717441/

[8] http://www.datanami.com/2013/07/19/big_data_emerges_in_indian_hea lth_care/

[9] http://eandt.theiet.org/magazine/2013/03/journey-to-the-centre-of-big- data.cfm

[10] http://itbussinessconsulting.jimdo.com/2014/08/11/how-big-data-and- analytics-can-reshape-healthcare/

[11] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare:promise and potential," *Health Information Science and Systems*, vol. 2, no. 3.